Selection of latent variables for multiple mixed-outcome models

LING ZHOU, HUAZHEN LIN
Center of Statistical Research, School of Statistics,
Southwestern University of Finance and Economics

XINYUAN SONG
Department of Statistics,
The Chinese University of Hong Kong

YI LI
Department of Biostatistics
University of Michigan

ABSTRACT. Latent variable models have been widely used for modeling the dependence structure of multiple outcomes data. However, the formulation of a latent variable model is often unknown *a priori*, the misspecification will distort the dependence structure and lead to unreliable model inference. Moreover, multiple outcomes with varying types present enormous analytical challenges. In this paper, we present a class of general latent variable models that can accommodate mixed types of outcomes. We propose a novel selection approach that simultaneously selects latent variables and estimates parameters. We show that the proposed estimator is consistent, asymptotically normal and has the oracle property. The practical utility of the methods is confirmed via simulations as well as an application to the analysis of the World Values Survey, a global research project that explores peoples' values and beliefs and the social and personal characteristics that might influence them.

Key words: dependence structure, latent variables model, oracle property, SCAD penalty, selection of latent variables

1 Introduction

Multiple outcomes that include both continuous and ordinal variables are often collected in applications where the responses of interest cannot be measured directly, or are difficult or expensive to measure. Latent variable models (LVMs) are commonly adopted, which state that different outcomes are conditionally independent measures of the latent variables, possibly capturing various aspects of them. Thus, unlike conventional random effects, which are mainly used to address the heterogeneity or dependence among observed outcomes, latent variables represent theoretical concepts or constructs that cannot be directly assessed by a single observed variable, but instead are measured through multiple observed variables. In practice, the formulation of an LVM (e.g., what and how many latent variables should be included) is often unknown *a priori*. Misspecification of the model would distort the dependence structure and lead to unreliable model inference (Leek & Storey, 2008). In particular, overspecified LVMs may result in highly correlated latent variables of which the covariance matrix becomes singular or nearly singular, leading to both theoretical and computational difficulties. Hence, a fundamental problem in the analysis of LVMs is model selection, especially the selection of latent variables that are relevant to substantive study.

The existing work on LVMs focuses on the estimation of model parameters; limited work has been devoted to the selection of latent variables, predominantly within the framework of factor analysis models—the most basic version of LVMs. For example, the Akaike information criterion (AIC; Akaike, 1987), Bayesian information criterion (BIC; Schwarz, 1978) and Bayesian approaches have been proposed to select the factors in factor analysis models (Press & Shigemasu, 1989 and 1997; Lee & Song, 2002; Carvalho *et al.*, 2005; Bhattacharya & Dunson, 2011). However, these methods incur a heavy computational burden and quickly become infeasible when the number of possible factors becomes even moderately large. In addition, the large sample model selection results (e.g., model selection consistency and oracle property) are elusive, making it difficult to evaluate the procedure's statistical properties.

We propose a new penalized pseudo-likelihood method that selects latent variables and estimates regression parameters simultaneously for a general LVM. Because the factor analysis model is a special case of the general LVM, our method can be used to select the factors in factor analysis models. However, different from existing work on factor selection in factor analysis models, our method reduces the computational

burden in that it does not require *a priori* specification of all possible latent variables. Furthermore, our estimator is shown to have desirable theoretical properties, including $n^{1/2}$-consistency, asymptotic normality and the oracle property—that is, it works as well as if the latent variables were known.

Though related, our context is different from that of random effect selection in random effect models. Indeed, random effects are mainly introduced to describe the unobserved heterogeneity and are covariate-independent, whereas latent variables represent specific traits associated with covariates and hence are covariate-dependent. As a result, the methods for selecting random effects cannot be applied to the selection of latent factors (Chen & Dunson, 2003). However, as described in Section 3, the proposed method can also be used to select random effects.

Analysis of multiple outcomes is further complicated by the fact that the outcomes can typically be of mixed types (i.e., binary, continuous or ordinal), which presents statistical challenges, as a natural multivariate distribution for mixed data does not exist. Yang *et al.* (2007) and Wagner & Tüchler (2010) considered joint models for Poisson and continuous data. Muthén (1984) proposed to define ordinal variables using unknown threshold parameters applied to underlying normal continuous variables. However, the literature on underlying normal models has focused primarily on joint models for low-dimensional ordinal outcomes and continuous outcomes (Catalano & Ryan, 1992; Cox & Wermuth, 1992; Fitzmaurice & Laird, 1995; Sammel *et al.*, 1997; Regan & Catalano, 1999; Dunson, 2000; Roy & Lin, 2000; Gueorguieva & Agresti, 2001). This paper proposes a two-step approach for jointly modeling continuous, binary and ordinal outcomes data under the underlying normal framework. Our estimation and selection procedure utilizes a closed-form penalized maximum likelihood estimator, which greatly facilitates computation.

The remainder of the paper is organized as follows. We introduce the proposed general LVM in Section 2. We propose a new penalized pseudo-likelihood method that allows us to select latent variables and estimate regression and threshold parameters simultaneously in Section 3. To implement the proposal, we provide a series of estimating equation-based approaches to draw inference and further propose a BIC-type procedure to select tuning parameters. In Section 4, we state our estimators' theoretical properties, including $n^{1/2}$-consistency, asymptotic normality and the oracle property. We report in Section 5 simulation results and an analysis of the World Values Survey (WVS), a global research project that explores the social and personal characteristics that influence people's values and beliefs. We provide concluding re-

marks in Section 6. We defer all proofs to the Supplementary Material.

2 General latent variable model

Suppose there are n randomly selected subjects, each with p distinct outcomes. Specifically, for the ith subject, we observe vectors of covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$, and a vector of outcomes $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{ip})'$, where the first element of $\mathbf{X}_i$ is 1. Without loss of generality, we assume that the first p_1 elements of $\mathbf{Y}_i$ are continuous and that the remaining $p_2 = p - p_1$ elements are ordinal and are linked to some underlying continuous variables as in Muthén (1984). That is, $Y_{ij} = g_j(U_{ij}; \mathbf{c}_j)$ for $j = 1, \cdots, p$, where U_{ij} is a continuous underlying variable of Y_{ij}. For the continuous outcomes, we have $Y_{ij} = U_{ij}$, for $j = 1, \cdots, p_1$; for an ordinal outcome $Y_{ij} \in \{1, \cdots, d_j\}$, where $d_j \geq 2$ is a positive integer, we have $Y_{ij} = \sum_{l=1}^{d_j} lI(c_{j,l-1} < U_{ij} \leq c_{j,l})$ for $j = p_1 + 1, \cdots, p$, where $\mathbf{c}_j = (c_{j,0}, \cdots, c_{j,d_j})'$ are thresholds satisfying $-\infty = c_{j,0} < c_{j,1} < \cdots < c_{j,d_j} = \infty$. In summary, $g_j(\cdot)$ is the identity link for continuous outcomes and is otherwise a threshold link mapping from $\mathbb{R} \to \{1, \cdots, d_j\}$ for the jth outcome. Let $\boldsymbol{\xi}_i = (\xi_{i1}, \cdots, \xi_{iq})'$, a q-dimensional random vector of latent variables that represents an individual's specific traits, $q \leq p$. We then relate the underlying continuous variables $\mathbf{U}_i = (U_{i1}, \cdots, U_{ip})'$ to $\boldsymbol{\xi}_i$ via

$$\mathbf{U}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\alpha}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)'$ is a regression coefficient matrix, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p)'$ is a loading matrix with vector $\boldsymbol{\alpha}_j = (\alpha_{j1}, \cdots, \alpha_{jq})'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{ip})'$ is a vector of random errors distributed as $N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$ with $\boldsymbol{\Sigma}_\varepsilon = \text{diag}(\sigma^2_{\varepsilon 1}, \cdots, \sigma^2_{\varepsilon p})$. Model (2.1) assumes that multiple outcomes are independent given latent variables, implying that the correlation among $Y_{ij}, j = 1, \cdots, p$ is due entirely to the shared latent variables in $\boldsymbol{\xi}_i$, explaining all the dependence among responses.

We stress that, unlike random effects, the latent variables $\boldsymbol{\xi}_i$ are introduced to reflect an individual's unobservable traits, such as 'life satisfaction' and 'job attitude', which, as in Roy & Lin (2000) and Skrondal & Rabe-Hesketh (2007), are linked to observed covariates via

$$\boldsymbol{\xi}_i = \boldsymbol{\gamma}\mathbf{Z}_i + \mathbf{e}_i, \tag{2.2}$$

where $\mathbf{e}_i = (e_{i1}, \cdots, e_{iq})' \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e)$ is a vector of random errors independent of $\mathbf{Z}_i$, and $\boldsymbol{\Sigma}_e = \text{diag}(\sigma^2_{e1}, \cdots, \sigma^2_{eq})$. Here, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_q)'$ is a matrix of unknown regression coefficients with vector $\boldsymbol{\gamma}_j = (\gamma_{j1}, \cdots, \gamma_{jm})'$ and is used to describe effects

of observed predictors on latent variables and then on outcomes. We term model (2.2), coupled with model (2.1), a general LVM as it extends the common LVM by accommodating both continuous and ordinal outcomes. The covariates in $\mathbf{X}_i$ and $\mathbf{Z}_i$ play different roles in the proposed model; $\mathbf{Z}_i$ records the covariates of interest and is used to characterize the latent variables, whereas $\mathbf{X}_i$ exclusive of $\mathbf{Z}_i$ is used to adjust for subjects' characteristics that may affect the outcomes. In model (2.2), the latent variable $\xi_{ij} = \boldsymbol{\gamma}_j'\mathbf{Z}_i + e_{ij}$ is zero if $\sigma_{ej} = 0$ and $\|\boldsymbol{\gamma}_j\| = 0$; ξ_{ij} is an observed covariate for $\sigma_{ej} = 0$ if only one $\gamma_{jk} \neq 0$ among $\{\gamma_{jk}, k = 1, \cdots, m\}$ and is a linear combination of observed covariates otherwise; ξ_{ij} is a random intercept if $\sigma_{ej} \neq 0$ and $\|\boldsymbol{\gamma}_j\| = 0$; ξ_{ij} is a latent variable if $\sigma_{ej} \neq 0$ and $\|\boldsymbol{\gamma}_j\| \neq 0$ (particularly when $\sigma_{ej} \neq 0$, $\gamma_{jk} \neq 0$ $(k \in \mathcal{A})$ and $\gamma_{jk} = 0$ $(k \notin \mathcal{A})$); and ξ_{ij} is a latent variable characterized by the predictors $\{Z_{ik}, k \in \mathcal{A}\}$. However, the latent variables or random effects to be included in models (2.1) and (2.2) are often unknown *a priori*, which presents a dilemma: too few latent variables would lead to a large modeling bias, whereas too many would result in overfitting. This inevitably leads to the task of selecting important latent variables. On the other hand, as model (2.2) stipulates, certain predictors influence the responses only through intermediate latent variables, meaning that latent variables are characterized by subsets of predictors $\mathbf{Z}_i$. In practice, identification of such subsets of latent variables is important in that it facilitates interpretation. Therefore, it is essential to develop a procedure that automatically selects latent variables and the corresponding underlying subsets of predictors.

To proceed, we first discuss the identifiability issue of models (2.1) and (2.2), which can be rewritten as

$$\mathbf{U}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\alpha}\mathbf{e}_i + \boldsymbol{\varepsilon}_i. \tag{2.3}$$

Hence $\mathbf{U}_i \sim N(\boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\alpha}\boldsymbol{\Sigma}_e\boldsymbol{\alpha}' + \boldsymbol{\Sigma}_\varepsilon$. Given that only $\boldsymbol{\alpha}\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}\boldsymbol{\Sigma}_e\boldsymbol{\alpha}'$ are identifiable, we follow the common practice in factor analysis (Anderson & Rubin, 1956; Lee, 2007; Lee & Song, 2002) to introduce the constraints $\alpha_{jk} = 0$ for all $j < k$, where $j = 1, \cdots, p$, $k = 1, \cdots, q \leq p$, to eliminate the indeterminacy of rotation in a model with q factors, and introduce constraints $\alpha_{kk} = 1$, $k = 1, \cdots, q$ to fix the sign of each column of $\boldsymbol{\alpha}$. To identify the ordinal variables, we further set $\sigma_{\epsilon j} = 1$ for $j > p_1$ (Dunson, 2000; Shi & Lee, 2000; Lee & Song, 2004) and exclude the intercept term from $\mathbf{X}_i$. This way, all $\boldsymbol{\alpha}, \boldsymbol{\Sigma}_e$ and $\boldsymbol{\gamma}$ are identifiable.

Although related, the proposed model (2.3) with regressors $(\mathbf{X}; \mathbf{Z})$ and the particular covariance error structure differs from an ordinary mixed effect model. The random effects in the latter address the heterogeneity or dependence of the data

but have no specific meaning, whereas the latent variables in model (2.3) represent certain unobservable traits that are characterized by some covariates. Thus, model (2.3) not only addresses the heterogeneity, but also provides insights into the causes and effects of such heterogeneity, consequently increasing its capability in terms of interpretation.

3 Selection and estimation

3.1. Penalized likelihood function

Let $\mathbf{U}_i = (\mathbf{U}_{i1}', \mathbf{U}_{i2}')'$, where $\mathbf{U}_{i1}$ corresponds to the first p_1 continuous components—which are completely observed—and $\mathbf{U}_{i2}$ is a collection of U_{ij} corresponding to the last $p - p_1$ discrete components. For example, $Y_{ij} = k$ implies that U_{ij} falls into $[c_{j,k-1}, c_{j,k})$, where $\{c_{j,k}\}$ are threshold parameters and need to be estimated. Let $\mathcal{A}_i = \prod_{j=p_1+1}^{p} [c_{j,Y_{ij}-1}, c_{j,Y_{ij}})$. Then the likelihood for the observed data $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_n\}$ can be expressed as

$$L_n(\boldsymbol{\Theta}) \propto \prod_{i=1}^{n} |\boldsymbol{\Sigma}|^{-1/2} \int_{\mathbf{U}_{i2} \in \mathcal{A}_i} \exp\left[-\frac{1}{2} \left\{ \begin{pmatrix} \mathbf{U}_{i1} \\ \mathbf{U}_{i2} \end{pmatrix} - \boldsymbol{\beta}\mathbf{X}_i - \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i \right\}' \times \boldsymbol{\Sigma}^{-1} \left(\begin{pmatrix} \mathbf{U}_{i1} \\ \mathbf{U}_{i2} \end{pmatrix} - \boldsymbol{\beta}\mathbf{X}_i - \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i \right) \right] d\mathbf{U}_{i2}, \tag{3.1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\varepsilon, \boldsymbol{\Sigma}_e, \boldsymbol{\gamma}\}$ includes all unknown structural parameters. We assume $\{c_{j,k}\}$ to be known for now, and we estimate them in Section 3.3.

As explained in Section 2, ξ_{ij} may be a latent variable, random effect, manifest variable (that is, observable variable) or zero, depending on whether σ_{ej} and $\|\boldsymbol{\gamma}_j\|$ are zero. If ξ_{ij} is a latent variable, it is of interest to know the corresponding subset of predictors. The selection of the subset corresponds to some elements of $\{\gamma_{jk}, k = 1, \ldots, m, j = 1, \ldots, q\}$ being zero, which leads to the following likelihood with penalties on $(\sigma_{ej}, \gamma_{jk}, k = 1, \ldots, m, j = 1, \ldots, q)'$,

$$Q(\boldsymbol{\Theta}) = \log L_n(\boldsymbol{\Theta}) - n \sum_{j=1}^{q} p_{\rho_{1n}}(\sigma_{ej}) - n \sum_{j=1}^{q} \sum_{k=1}^{m} p_{\rho_{2n}}(|\gamma_{jk}|). \tag{3.2}$$

Here, $p_\lambda(\cdot)$ is a penalty function, the common choices of which include L_q penalty, $p_\lambda(|\beta|) = \lambda|\beta|^q, (q > 0)$, yielding the well-known ridge regression with $q = 2$. The smoothly clipped absolute deviation (SCAD) penalty function (Fan & Li, 2001) takes

the form

$$\dot{p}_\lambda(\beta) = \lambda \left\{ I(\beta \leqslant \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \text{ for some } a > 2 \text{ and } \beta > 0, \quad (3.3)$$

with $\dot{p}_\lambda(0) = 0$, where $\dot{f}(t) = df(t)/dt$ for any smooth function f. The tuning parameter a is often taken to be 3.7 as suggested by Fan & Li (2001). As the SCAD penalty has been shown to render oracle properties in many penalized likelihood settings (Fan *et al.*, 2006), we adopt it in our ensuing development. However, our method does accommodate more general penalty functions.

Indeed, by maximizing the penalized likelihood $Q(\boldsymbol{\Theta})$, we can show that there is a positive probability of some estimated values of σ_{ej} and γ_{jk} equaling zero and thus of automatically selecting latent variables and corresponding predictors. Thus, the procedure combines the selection of latent variables and corresponding subsets of predictors, with the estimation of parameters into one step, reducing the computational burden substantially.

3.2. Penalized expectation maximization algorithm

With the likelihood function $L_n(\boldsymbol{\Theta})$ involving a $p - p_1$ dimensional intractable integral, a direct application of the maximum likelihood (ML) estimation procedure is nearly impossible. We propose below a penalized expectation maximization (EM) algorithm. Given the complexity of the proposed algorithm, we describe the basic steps and computation of the conditional means required for the maximization in two subsections.

3.2.1. The basic steps of the penalized EM algorithm

The random variable $e_{ij} \sim N(0, \sigma_{ej}^2)$ if $\sigma_{ej} \neq 0$; otherwise, $e_{ij} \equiv 0$. Hence, $\mathbf{e}_i$ is a mixture of zero and normal components. For ease of presentation, we rewrite $\mathbf{e}_i = \boldsymbol{\Sigma}_e^{1/2} \mathbf{w}_i$, where $\mathbf{w}_i = (w_{i1}, \cdots, w_{iq})' \sim N(0, I)$. Then, model (2.3) can be rewritten as

$$\mathbf{U}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\alpha}\boldsymbol{\Sigma}_e^{1/2}\mathbf{w}_i + \boldsymbol{\varepsilon}_i. \quad (3.4)$$

To set up a penalized EM algorithm, consider the random variables $\mathbf{U}_{i2}$ and $\mathbf{w}_i$ to be the missing data. The complete data for individual i is $D_i = \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{U}_i, \mathbf{w}_i\}$. The penalized complete-data log-likelihood function is

$$\mathcal{Q}_c(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}) - n \sum_{j=1}^{q} p_{\rho_{1n}}(\sigma_{ej}) - n \sum_{j=1}^{q} \sum_{k=1}^{m} p_{\rho_{2n}}(|\gamma_{jk}|), \quad (3.5)$$

where

$$\log L(\boldsymbol{\Theta}) \propto -\frac{1}{2}\sum_{i=1}^{n}\left[\sum_{j=1}^{p}\left\{\log\sigma_{\varepsilon j}^2+\frac{(U_{ij}-\mathbf{X}_i'\boldsymbol{\beta}_j-\boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i-\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e^{1/2}\mathbf{w}_i)^2}{\sigma_{\varepsilon j}^2}\right\}\right]. \quad (3.6)$$

In the maximization step, we maximize the conditional expectation of $\mathcal{Q}_c(\boldsymbol{\Theta})$ given the observed data. The maximization step depends on the conditional expectation of some function of $\mathbf{U}_{i2}$ and $\mathbf{w}_i$, which is evaluated in the expectation step. The two steps are iterated until convergence.

3.2.2. Implementation of the penalized EM algorithm

Let $\delta_{ij}(\boldsymbol{\Theta}) = U_{ij}-\mathbf{X}_i'\boldsymbol{\beta}_j-\boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i-\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e^{1/2}\mathbf{w}_i$. For any given threshold parameter $c_{j,k}$, we estimate $\boldsymbol{\Theta}$ by maximizing $E\{\mathcal{Q}_c(\boldsymbol{\Theta})|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i,i=1,\cdots,n\}$ with respect to $\boldsymbol{\Theta}$. Differentiating $E\{\mathcal{Q}_c(\boldsymbol{\Theta})|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i,i=1,\cdots,n\}$ with respect to $\boldsymbol{\Theta}$ and setting the derivatives to zero leads to the following estimation equations:

$$\sigma_{\varepsilon j}^2=\frac{1}{n}\sum_{i=1}^{n}E\left\{\delta_{ij}(\boldsymbol{\Theta})^2|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\right\} \text{ for } j=1,\cdots,p, \quad (3.7)$$

$$\boldsymbol{\beta}_j=\left(\sum_{i=1}^{n}\frac{\mathbf{X}_i\mathbf{X}_i'}{\sigma_{\varepsilon j}^2}\right)^{-1}\sum_{i=1}^{n}\frac{\mathbf{X}_iE(U_{ij}-\boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i-\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e^{1/2}\mathbf{w}_i|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i)}{\sigma_{\varepsilon j}^2} \text{ for } j=1,\cdots,p, \quad (3.8)$$

$$\alpha_{jk}=\left[\sum_{i=1}^{n}\frac{E\left\{(\mathbf{Z}_i'\boldsymbol{\gamma}_k+\sigma_{ek}w_{ik})^2|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\right\}}{\sigma_{\varepsilon j}^2}\right]^{-1}$$
$$\times\left[\sum_{i=1}^{n}\frac{E\left\{\left(U_{ij}-\mathbf{X}_i'\boldsymbol{\beta}_j-\sum_{m\neq k}\alpha_{jm}(\boldsymbol{\gamma}_m'\mathbf{Z}_i+\sigma_{em}w_{im})\right)(\mathbf{Z}_i'\boldsymbol{\gamma}_k+\sigma_{ek}w_{ik})\,|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\right\}}{\sigma_{\varepsilon j}^2}\right],$$
$$\text{for } j=1,\cdots,p \text{ and } k<j, \quad (3.9)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{p}\frac{\alpha_{kj}Z_{ir}E\left\{\delta_{ik}(\boldsymbol{\Theta})|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\right\}}{\sigma_{\varepsilon k}^2}-n\dot{p}_{\rho_{2n}}(|\gamma_{jr}|)sgn(\gamma_{jr})=0,$$
$$\text{for } j=1,\cdots,q, r=1,\cdots,m, \quad (3.10)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{p}\frac{\alpha_{kj}E(w_{ij}\delta_{ik}(\boldsymbol{\Theta})|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i)}{\sigma_{\varepsilon k}^2}-n\dot{p}_{\rho_{1n}}(\sigma_{ej})=0 \text{ for } j=1,\cdots,q. \quad (3.11)$$

We estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}_e$ by rewriting equations (3.10) and (3.11) as

$$\gamma_{jr}=\left(\sum_{i=1}^{n}\sum_{k=1}^{p}\frac{\alpha_{kj}^2Z_{ir}^2}{\sigma_{\varepsilon k}^2}+n\dot{p}_{\rho_{2n}}(|\gamma_{jr}|)/|\gamma_{jr}|\right)^{-1}$$
$$\times\sum_{i=1}^{n}\sum_{k=1}^{p}\frac{\alpha_{kj}Z_{ir}}{\sigma_{\varepsilon k}^2}E\left(U_{ik}-\mathbf{X}_i'\boldsymbol{\beta}_k-\sum_{m\neq j}\alpha_{km}\boldsymbol{\gamma}_m'\mathbf{Z}_i-\sum_{l\neq r}\alpha_{kj}Z_{il}\gamma_{jl}-\boldsymbol{\alpha}_k'\boldsymbol{\Sigma}_e^{1/2}\mathbf{w}_i|\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i\right),$$
$$\text{for } j=1,\cdots,q, r=1,\cdots,m, \quad (3.12)$$

and

$$\sigma_{ej} = \left\{ \sum_{i=1}^{n} \sum_{k=1}^{p} \frac{\alpha_{kj}^2 E(w_{ij}^2 | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)}{\sigma_{\varepsilon k}^2} + n\dot{p}_{\rho_{1n}}(\sigma_{ej})/\sigma_{ej} \right\}^{-1}$$
$$\times \left\{ \sum_{i=1}^{n} \sum_{k=1}^{p} \frac{\alpha_{kj} E\left[w_{ij} \left(U_{ik} - \mathbf{X}_i'\boldsymbol{\beta}_k - \boldsymbol{\alpha}_k'\boldsymbol{\gamma}\mathbf{Z}_i - \sum_{m \neq j} \alpha_{km}\sigma_{em} w_{im} \right) | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i \right]}{\sigma_{\varepsilon k}^2} \right\},$$
$$\text{for } j = 1, \cdots, q. \qquad (3.13)$$

Then, we estimate $\boldsymbol{\Theta}$ by repeatedly using equations (3.7), (3.8), (3.9), (3.12) and (3.13) until $\boldsymbol{\Theta}$ converges. For each step, $\boldsymbol{\Theta}$ in the left side of the equations is replaced by the value from the last step.

To obtain the estimate of $\boldsymbol{\Theta}$ using the above equations, we need to compute the conditional mean and conditional variance matrices of $(\mathbf{U}_{i2}, \mathbf{w}_i)$ given $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, which has the form of $E\big(\mathbf{U}_{i2}^{\otimes r_1} \otimes \mathbf{w}_i^{\otimes r_2} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\big)$ for $r_1 + r_2 \leq 2$, $r_1 = 0, 1, 2$, and $r_2 = 0, 1, 2$, where $a^{\otimes 2} = aa'$, $a^{\otimes 1} = a$ and $a \otimes b = ab'$. Because $E\left(\mathbf{U}_{i2}^{\otimes r_1} \otimes \mathbf{w}_i^{\otimes r_2} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right) = E\left\{\mathbf{U}_{i2}^{\otimes r_1} \otimes E\left(\mathbf{w}_i^{\otimes r_2} | \mathbf{U}_i, \mathbf{X}_i, \mathbf{Z}_i\right) | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right\}$, and given $\mathbf{U}_i, \mathbf{X}_i$ and $\mathbf{Z}_i$, $\mathbf{w}_i$ is a normal random variable with mean $\boldsymbol{\Sigma}_e^{1/2}\boldsymbol{\alpha}' \left(\boldsymbol{\alpha}\boldsymbol{\Sigma}_e\boldsymbol{\alpha}' + \boldsymbol{\Sigma}_\varepsilon\right)^{-1} \left(\mathbf{U}_i - \boldsymbol{\beta}\mathbf{X}_i - \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i\right)$ and covariance matrix $I - \boldsymbol{\Sigma}_e^{1/2}\boldsymbol{\alpha}' \left(\boldsymbol{\alpha}\boldsymbol{\Sigma}_e\boldsymbol{\alpha}' + \boldsymbol{\Sigma}_\varepsilon\right)^{-1} \boldsymbol{\alpha}\boldsymbol{\Sigma}_e^{1/2}$. To calculate $E\left(\mathbf{U}_{i2}^{\otimes r_1} \otimes \mathbf{e}_i^{\otimes r_2} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right)$, it is sufficient to compute $E\left(\mathbf{U}_{i2}^{\otimes r} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right)$, for $r = 1, 2$, which is

$$E\left(\mathbf{U}_{i2}^{\otimes r} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right) = E\left\{\mathbf{U}_{i2}^{\otimes r} I\left(\mathbf{U}_{i2} \in \mathcal{A}_i\right) | \mathbf{U}_{i1}, \mathbf{X}_i, \mathbf{Z}_i\right\} / P\left(\mathbf{U}_{i2} \in \mathcal{A}_i | \mathbf{U}_{i1}, \mathbf{X}_i, \mathbf{Z}_i\right),$$

where both the numerator and denominator can be approximated with Monte Carlo simulations.

3.3. Estimation of the threshold parameters

We are now in a position to estimate $\{c_{j,k}\}$ with the iterative series of estimating equations proposed below. The parameters $\boldsymbol{\Theta}$ are then updated by maximizing the pseudo-likelihood $E\{\mathcal{Q}_c | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, i = 1, \cdots, n\}$, with $\{c_{j,k}\}$ replaced by their estimated values. The procedure is repeated until convergence.

Because $U_{ij} = \mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\alpha}_j'\mathbf{e}_i + \varepsilon_{ij}$, for any given $j > p_1$, $k \in \{1, \cdots, d_j\}$, $\mathbf{X}_i$, and $\mathbf{Z}_i$, we have

$$Pr(Y_{ij} = k | \mathbf{X}_i, \mathbf{Z}_i) = \Phi\left\{ \frac{c_{j,k} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\} - \Phi\left\{ \frac{c_{j,k-1} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random

variable. With $c_{j,0} = -\infty$, we estimate $c_{j,1}, \cdots, c_{j,d_j-1}$, one-at-a-time, using

$$\sum_{i=1}^{n} \left[I\left(Y_{ij} = k\right) - \Phi\left\{ \frac{c_{j,k} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\} + \Phi\left\{ \frac{c_{j,k-1} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\boldsymbol{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\} \right] = 0, \tag{3.14}$$

for $k = 1, \cdots, d_j - 1$.

3.4. Selection of tuning parameters

We select the tuning parameters ρ_{1n} and ρ_{2n} using a BIC-based procedure. As shown by Wang *et al.* (2007), such a procedure typically yields model selection consistency for linear regression models. Specifically, we choose ρ_{1n} and ρ_{2n} separately as they control the complexity of two separate components of models. First, noting that ρ_{2n} controls the number of non-zero elements in $\boldsymbol{\gamma}$, we rewrite model (2.3) as

$$\mathbf{U}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\alpha}\boldsymbol{\gamma}\mathbf{Z}_i + \tilde{\boldsymbol{\varepsilon}}_i, \tag{3.15}$$

where $\tilde{\boldsymbol{\varepsilon}}_i = \boldsymbol{\alpha}\mathbf{e}_i + \boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma})$. The parameters $\boldsymbol{\gamma}$ are regression coefficients. We then select the optimal ρ_{2n} by maximizing

$$BIC_2 = \log L_n(\boldsymbol{\Theta}) - \frac{1}{2} DF_{\rho_{2n}} log(np), \tag{3.16}$$

where $L_n(\boldsymbol{\Theta})$ is the observed-data likelihood function defined by model (3.1) and $DF_{\rho_{2n}}$ is the generalized degree of freedom, which can be consistently estimated by $\sum_{j=1}^{q}\sum_{k=1}^{m} I(|\widehat{\gamma}_{jk}| \neq 0) + \sum_{j=1}^{p}\sum_{k=1}^{q} I(\widehat{\alpha}_{jk} \neq 0) + \sum_{j=1}^{q} I(\widehat{\sigma}_{ej} \neq 0)$, the number of nonzero coefficients; see Zhang *et al.* (2010) for models with generalized linear structure.

We now discuss choice of ρ_{1n}, which controls the dimension of the random effect $\mathbf{e}_i$—that is, the number of non-zero elements in $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, \cdots, \sigma_{eq}^2)$. Model (3.4) shows that $\boldsymbol{\Sigma}_e^{1/2}$ is the regression effect of $\mathbf{w}_i$. To select $\boldsymbol{\Sigma}_e$, we thus consider the random variable $\mathbf{w}_i$ and the covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$ as input variables in model (2.3) and only $\boldsymbol{\varepsilon}_i$ as random noise. We then select the optimal ρ_{1n} by maximizing

$$BIC_1 = E\{\log L(\boldsymbol{\Theta}) | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, i = 1, \cdots, n\} - \frac{1}{2} DF_{\rho_{1n}} log(np), \tag{3.17}$$

where $L(\boldsymbol{\Theta})$ is the complete-data likelihood function defined by equation (3.6), $DF_{\rho_{1n}}$ is the weighted generalized degree of freedom $DF_{\rho_{1n}} = \sum_{i=1}^{q} w_i I(\widehat{\sigma}_{ei} \neq 0) + \sum_{j=1}^{q}\sum_{k=1}^{m} I(|\widehat{\gamma}_{jk}| \neq 0) + \sum_{j=1}^{p}\sum_{k=1}^{q} I(\widehat{\alpha}_{jk} \neq 0)$ with $w_i = 1/\widehat{\sigma}_{ei}^{ini}$, and $\widehat{\sigma}_{ei}^{ini}$ is the estimate of σ_{ei} without

penalty. Here, we replace the complete data likelihood with the conditional expectation of the complete data likelihood, because the complete data likelihood depends on the missing data $\mathbf{w}_i$ and is useless in the estimation of ρ_{1n}. However, the conditional expectation of the complete data likelihood is a reasonable estimator for the complete data likelihood. We test the performance of our tuning procedure via simulation studies in Section 5. In simulation studies and in the analysis of actual data, we perform the selection of both ρ_{1n} and ρ_{2n} on grids of the tuning parameters.

4 Large sample properties

We now establish the consistency and asymptotic normality of the proposed estimator. For ease of presentation, we rewrite $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \boldsymbol{\sigma}_e{}', \vec{\boldsymbol{\gamma}}')'$ as the vectorial form of the collection of all unknown parameters. Here $\boldsymbol{\Theta}_1 = (\vec{\boldsymbol{\alpha}}', \vec{\boldsymbol{\beta}}', \boldsymbol{\sigma}_\varepsilon{}')'$. Throughout, we use the subscript "0" to represent the true value. Without loss of generality, let $\boldsymbol{\sigma}_{e0} = (\boldsymbol{\sigma}'_{e(1)0}, \boldsymbol{\sigma}'_{e(2)0})'$, $\vec{\boldsymbol{\gamma}}_0 = (\boldsymbol{\gamma}'_{(1)0}, \boldsymbol{\gamma}'_{(2)0})'$ and $\boldsymbol{\sigma}_{e(2)0} = 0$ and $\boldsymbol{\gamma}_{(2)0} = 0$. Define $\boldsymbol{\sigma}_e = (\boldsymbol{\sigma}'_{e(1)}, \boldsymbol{\sigma}'_{e(2)})'$, $\vec{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}'_{(1)}, \boldsymbol{\gamma}'_{(2)})'$ to have the corresponding decompositions.

Considering a more generalized nonconcave penalty function, we set $a_{n1} = \max_j\{\dot{p}_{\rho_{1n}}(\sigma_{ej0}) : \sigma_{ej0} \neq 0\}$, $a_{n2} = \max_{j,k}\{\dot{p}_{\rho_{2n}}(|\gamma_{jk0}|) : |\gamma_{jk0}| \neq 0\}$ and $a_n = \max\{a_{n1}, a_{n2}\}$. Let $\ddot{g}(t) = d^2g(t)/dt^2$. The following theorems summarize the large sample properties of the proposed estimator; their proofs are deferred to the Supplementary Material, and the related regularity conditions are given in Appendix A.2.

Theorem 1 *Under conditions 1–3 stated in Appendix A.2, if* $\max_j\{|\ddot{p}_{\rho_{1n}}(\sigma_{ej0})| : \sigma_{ej0} \neq 0\} \to 0$ *and* $\max_{j,k}\{|\ddot{p}_{\rho_{2n}}(|\gamma_{jk0}|)| : |\gamma_{jk0}| \neq 0\} \to 0$*, then, as* $n \to \infty$*,*

(1) for any $j = p_1 + 1, \cdots, p$, $k \in \{1, \cdots, d_j\}$*, we have*

$$\widehat{c}_{j,k} \to_P c_{j,k0} \quad \textit{and} \quad \|\widehat{c}_{j,k} - c_{j,k0}\| = O_p(n^{-1/2} + a_n). \tag{4.1}$$

(2) there is a maximizer $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Theta}}_1', \widehat{\boldsymbol{\sigma}}_e', \widehat{\vec{\boldsymbol{\gamma}}}')'$ *of* $Q(\boldsymbol{\Theta})$ *such that*

$$\|\widehat{\boldsymbol{\sigma}}_e - \boldsymbol{\sigma}_{e0}\| = O_p(n^{-1/2} + a_{n1}), \quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(n^{-1/2} + a_{n2}),$$
$$\textit{and} \quad \|\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{10}\| = O_p(n^{-1/2}). \tag{4.2}$$

Clearly, using the SCAD penalty defined in equation (3.3) with $\lambda \to 0$ and $\beta > 0$, we have $\dot{p}_\lambda(\beta) = \lambda\left\{\frac{(a\lambda-\beta)_+}{(a-1)\lambda}\right\} = \frac{(a\lambda-\beta)_+}{(a-1)} = 0$. Hence, with $\lambda = \rho_{1n} \to 0$ and $\lambda =$

$\rho_{2n} \to 0$, we obtain $a_{n1} = 0$ and $a_{n2} = 0$, respectively. Therefore, there exists a root-n consistent penalized estimator for the parameters $\mathbf{\Theta}$ and the threshold parameters $\mathbf{c}$. Next, we show that the penalized estimator demonstrates the oracle property.

Theorem 2 *Assume that the penalty function, $p_{\rho_{1n}}(\theta)$ and $p_{\rho_{2n}}(\theta)$, satisfies*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} \dot{p}_{\rho_{1n}}(\theta)/\rho_{1n} > 0, \text{ and } \liminf_{n\to\infty} \liminf_{\theta\to 0+} \dot{p}_{\rho_{2n}}(\theta)/\rho_{2n} > 0.$$

Under conditions 1–3 in Appendix A.2, if as $n \to \infty$, $\rho_{1n} \to 0$, $\sqrt{n}\rho_{1n} \to \infty$, $\rho_{2n} \to 0$ and $\sqrt{n}\rho_{2n} \to \infty$, the root-n consistent local maximizers $\widehat{\boldsymbol{\sigma}}_e = (\widehat{\boldsymbol{\sigma}}'_{e(1)}, \widehat{\boldsymbol{\sigma}}'_{e(2)})'$ and $\widehat{\vec{\boldsymbol{\gamma}}} = (\widehat{\vec{\boldsymbol{\gamma}}}_{(1)}, \widehat{\vec{\boldsymbol{\gamma}}}_{(2)})'$ in Theorem 1 must satisfy the following properties:

(a) Sparsity: $\widehat{\boldsymbol{\sigma}}_{e(2)} = 0$ and $\widehat{\vec{\boldsymbol{\gamma}}}_{(2)} = 0$.

(b) Asymptotic normality:

$$\sqrt{n}\left(\mathbf{\Lambda}_2 + \mathcal{U}_1\right)\left\{\widehat{\boldsymbol{\sigma}}_{e(1)} - \boldsymbol{\sigma}_{e(1)0} + \left(\mathbf{\Lambda}_2 + \mathcal{U}_1\right)^{-1}\left(C_{21}\mathbf{b}_1 + C_{22}\mathbf{b}_2\right)\right\} \to N\left(0, A_2\right) \text{ and}$$

$$\sqrt{n}\left(\mathbf{\Lambda}_3 + \mathcal{U}_2\right)\left\{\widehat{\vec{\boldsymbol{\gamma}}}_{(1)} - \vec{\boldsymbol{\gamma}}_{(1)0} + \left(\mathbf{\Lambda}_3 + \mathcal{U}_2\right)^{-1}\left(C_{31}\mathbf{b}_1 + C_{32}\mathbf{b}_2\right)\right\} \to N\left(0, A_3\right),$$

where $\mathbf{\Lambda}_2, \mathbf{\Lambda}_3, \mathcal{U}_1, \mathcal{U}_2, \mathbf{b}_1, \mathbf{b}_2, C_{21}, C_{22}, C_{31}, C_{32}, A_2$ and A_3 are defined in Appendix A.1.

Theorem 3 *When $n \to \infty$, if all conditions of Theorem 2 are satisfied, we have*

$$\sqrt{n}\mathbf{\Lambda}_1\left\{\widehat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_{10} + \mathbf{\Lambda}_1^{-1}\left(C_{11}\mathbf{b}_1 + C_{12}\mathbf{b}_2\right)\right\} \to N\left(0, A_1\right),$$

where $\mathbf{\Lambda}_1$, C_{11}, C_{12} and A_1 are defined in Appendix A.1.

Theorem 4 *When $n \to \infty$, if satisfying all the conditions of Theorem 2, we have*

$$\sqrt{n}\left\{\widehat{c}_{j,k} - c_{j,k0} + C_{4j1}(k)\mathbf{b}_1 + C_{4j2}(k)\mathbf{b}_2\right\} \to N\left\{0, A_{4j}(k)\right\},$$

where $C_{4j1}(k), C_{4j2}(k)$, and $A_{4j}(k)$ are defined in Appendix A.1.

For the SCAD penalty function, if $\rho_{1n} \to 0$ and $\rho_{2n} \to 0$, then $a_{n1} = a_{n2} = 0$, $\mathbf{b}_1 = 0$, $\mathbf{b}_2 = 0$, $\mathcal{U}_1 = 0$ and $\mathcal{U}_2 = 0$. Theorems 2–4 imply that the SCAD-based penalized likelihood estimators for $\boldsymbol{\sigma}_e$, $\boldsymbol{\gamma}$, Θ_1 and $c_{j,k}$ have the oracle property—that is, when the true parameters contain zero components, they are estimated as 0, with

the probability approaching 1, and the nonzero components are estimated as well as in the case where zero components are known.

In practice, to approximate the distribution and construct the confidence interval for $\widehat{\mathbf{\Theta}}_{(1)} = (\widehat{\mathbf{\Theta}}_1', \widehat{\boldsymbol{\sigma}}_{e(1)}', \widehat{\vec{\gamma}}_{(1)}')'$, the estimators of non-zero parameters, we need to estimate the variances of $\widehat{\mathbf{\Theta}}_{(1)}$. However, the complex form of the limiting covariance matrix of $\widehat{\mathbf{\Theta}}_{(1)}$ in Theorems 2 and 3 prohibits direct use. Instead, we propose using the resampling method of Jin *et al.* (2001) to estimate the variance. First, we generate n exponential random variables $V_i, i = 1, \cdots, n$ with mean 1 and variance 1. Then, we solve the following V_i-weighted estimation equations and denote the solutions as $\mathbf{\Theta}^*_{(1)}$ and $\mathbf{c}^*$:

$$\sum_{i=1}^{n} V_i \frac{\partial \log\{L_i(\mathbf{\Theta}; \mathbf{c})\}}{\partial \mathbf{\Theta}_{(1)}} |_{\boldsymbol{\sigma}_{e(2)}=0, \vec{\gamma}_{(2)}=0} = 0 \text{ and}$$

$$\sum_{i=1}^{n} V_i \left[I\left(Y_{ij} = k\right) - \Phi\left\{ \frac{c_{j,k} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\mathbf{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\} \right.$$
$$\left. + \Phi\left\{ \frac{c_{j,k-1} - \left(\mathbf{X}_i'\boldsymbol{\beta}_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i\right)}{\sqrt{\boldsymbol{\alpha}_j'\mathbf{\Sigma}_e\boldsymbol{\alpha}_j + 1}} \right\} \right] |_{\boldsymbol{\sigma}_{e(2)}=0, \vec{\gamma}_{(2)}=0} = 0, \quad \text{for} \quad \begin{matrix} k = 1, \cdots, d_j - 1 \\ j = p_1 + 1, \cdots, p \end{matrix},$$

with $c_{j,0} = -\infty$, where $L_i(\mathbf{\Theta}; \mathbf{c})$ is the observed-data likelihood function (3.1) for subject i. The estimates $\mathbf{\Theta}^*_{(1)}$ and $\mathbf{c}^*$ can be obtained using the same algorithm proposed in Sections 3.1−3.3. Using Theorems 2−4, the validity of the proposed resampling method is established as the following theorem. We omit its proof, as the arguments follow Jin *et al.* (2001).

Theorem 5 *Under the conditions of Theorem 2, the conditional distribution of* $n^{1/2}(\mathbf{\Theta}^*_{(1)} - \widehat{\mathbf{\Theta}}_{(1)})$ *given the observed data converges almost exactly to the asymptotic distribution of* $n^{1/2}(\widehat{\mathbf{\Theta}}_{(1)} - \mathbf{\Theta}_{(10)})$, *where* $\mathbf{\Theta}_{(10)}$ *is the true value of* $\mathbf{\Theta}_{(1)} = (\mathbf{\Theta}_1', \boldsymbol{\sigma}_{e(1)}', \vec{\gamma}_{(1)}')'$.

By repeatedly generating $V_1, \cdots, V_n$, we obtain a large number of realizations of $\mathbf{\Theta}^*_{(1)}$. The variance estimate of $\widehat{\mathbf{\Theta}}_{(1)}$ can be approximated by the empirical variance of $\mathbf{\Theta}^*_{(1)}$.

5 Simulation study

We have conducted extensive simulations to investigate the effect of misspecifying latent variables on the mean and the variance structure. Specifically, we consider the

model with two latent variables, denoted as LV2. In practice, the model selection procedure might reduce a latent variable to a manifest variable or a random effect. We hence compare the estimates from the proposed method with those from the following misspecified models: (1) the LV1MV1, where the variance of one latent variable is misspecified to 0–that is, one of latent variables is misspecified as a manifest variable; (2) the LV1RV1, where the regression coefficients of one latent variable are misspecified to 0–that is, one of the latent variables is misspecified as random effect.

We simulated 1000 data sets, each with $n = 200$ observations. For each subject, the latent variable is generated by the model $\xi_{ij} = \mathbf{Z}_i'\boldsymbol{\gamma}_j + e_{ij},\ \ j = 1, 2$, where $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3})'$, $Z_{ij}, j = 1, 2, 3$ are independently drawn from a standard normal random variable, $\boldsymbol{\gamma}_1 = (2, 0, 0)'$, $\boldsymbol{\gamma}_2 = (0, 2, 0)'$, $\mathbf{e}_i = (e_{i1}, e_{i2})'$ is a normal random vector with mean zero, and the covariance $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, \sigma_{e2}^2) = \text{diag}(1, 1)$. $\mathbf{e}_i$ and $\mathbf{Z}_i$ are independent. The outcomes $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ are generated from the models $Y_{ij} = X_{ij}'\boldsymbol{\beta}_j + \alpha_{j1}\xi_{i1} + \alpha_{j2}\xi_{i2} + \varepsilon_{ij},\ \ j = 1, 2, 3, 4$, where $\boldsymbol{\beta_1} = (\beta_{11}, \beta_{12})' = (1, 2)'$, $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22})' = (2, 2)'$, $\boldsymbol{\beta}_3 = (\beta_{31}, \beta_{32})' = (1, 1)'$, $\boldsymbol{\beta}_4 = (\beta_{41}, \beta_{42})' = (1.5, 2)'$, $X_{ij} = (1, X_{ij2})'$, and X_{ij2} is independently generated from a standard normal variable. Note that $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ are normal random vectors with mean zero and covariance $\boldsymbol{\Sigma}_\varepsilon \equiv \text{diag}(\sigma_{\varepsilon 1}^2, \sigma_{\varepsilon 2}^2, \sigma_{\varepsilon 3}^2, \sigma_{\varepsilon 4}^2) = \text{diag}(1, 1, 1, 1)$. $\boldsymbol{\alpha}' \equiv \begin{pmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ \alpha_{12} & \alpha_{22} & \alpha_{32} & \alpha_{42} \end{pmatrix} = \begin{pmatrix} 1 & 0.8 & 0.8 & 0.8 \\ 0 & 1 & 0.8 & 0.8 \end{pmatrix}$. For each simulated data set, we fit data with the LV2, LV1MV1 and LV1RV1 models and estimate the related unknown parameters using the ML method. The bias and empirical SDs of the estimators are reported in Table 1, where $\#CF$ is the number of convergence failures from 1000 simulation runs.

Using the data presented in Table 1, we make the following conclusions. (1) The estimate of the fixed effect in the measurement models are reported in the first part of Table 1. All estimators are unbiased, and LV2 has the smallest variance. The first part of Table 1 shows that misspecification of latent variables will lead to a slight loss of efficiency for $\boldsymbol{\beta}$. Misspecification of latent variables has a relatively minor effect on the parameters in the mean part. (2) The second part of Table 1 displays estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. A useful rule to keep in mind when checking bias, as suggested by Olsen & Schafer (2001), is that biases do not have a substantial negative effect on inference unless standardized bias (bias over SD) exceeds 0.4. By this rule, LV2 is unbiased, and LV1MV1 and LV1RV1 are seriously biased. Table 2 in the supplementary material shows that misspecification of latent variables leads to biased estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$,

the regression coefficients of the latent variable. (3) The third part of Table 1 shows the estimators of variances in the measurement and latent variable models. As shown, LV2 is unbiased and has the smallest variance; LV1RV1 and LV1MV1 are biased for the variance parameters in both the measurement and latent variable models.

Table 1: Estimation results for Simulation 1.

	$LV2(true)$	$LV1MV1$	$LV1RV1$
#CF	0	2	33
	Bias(SD)	Bias(SD)	Bias(SD)
β_{11}	-0.003(0.097)	-0.003(0.097)	-0.006(0.133)
β_{12}	-0.002(0.102)	-0.002(0.102)	0.001(0.139)
β_{21}	0.001(0.112)	0.002(0.113)	0.005(0.119)
β_{22}	-0.003(0.116)	-0.003(0.117)	-0.004(0.123)
β_{31}	0.000(0.106)	0.000(0.107)	0.002(0.109)
β_{32}	-0.002(0.106)	-0.002(0.107)	-0.003(0.109)
β_{41}	0.001(0.104)	0.001(0.105)	0.003(0.107)
β_{42}	0.000(0.107)	0.000(0.108)	-0.001(0.110)
α_{21}	-0.000(0.054)	0.940(0.248)	0.758(0.278)
α_{31}	0.000(0.050)	0.758(0.200)	0.618(0.229)
α_{41}	-0.000(0.049)	0.754(0.202)	0.613(0.226)
α_{32}	0.000(0.045)	0.003(0.046)	-0.085(0.291)
α_{42}	-0.003(0.043)	-0.000(0.044)	-0.076(0.300)
γ_{11}	-0.000(0.094)	-0.010(0.098)	-0.719(0.262)
γ_{12}	0.006(0.099)	0.004(0.101)	0.950(0.102)
γ_{13}	0(0)	0.002(0.101)	0.003(0.066)
γ_{21}	0(0)	-1.868(0.509)	*
γ_{22}	-0.001(0.111)	-0.009(0.172)	*
γ_{23}	0.002(0.097)	-0.003(0.168)	*
$\sigma^2_{\varepsilon 1}$	-0.030(0.173)	0.471(0.157)	2.119(0.472)
$\sigma^2_{\varepsilon 2}$	-0.028(0.153)	0.035(0.152)	-0.007(0.225)
$\sigma^2_{\varepsilon 3}$	-0.021(0.127)	-0.031(0.127)	-0.029(0.147)
$\sigma^2_{\varepsilon 4}$	-0.023(0.131)	-0.033(0.131)	-0.033(0.157)
σ^2_{e1}	-0.017(0.180)	-0.466(0.138)	-0.760(0.149)
σ^2_{e2}	-0.022(0.207)	*	0.472(0.809)

$*$ not applicable.

In summary, misspecification of latent variables has a minor effect on the estima-

tors of the parameters in the mean structure but may lead to biased estimators of the components of the covariance structure, including $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and the variances of the error and the latent variables.

As reported in the supplementary material, we have conducted further simulation studies (denoted as Simulation 2) to assess the finite-sample performance of the proposed method in terms of bias and empirical SD. We also examine the performance of criteria (3.16) and (3.17) in selecting ρ_{1n} and ρ_{2n}. We have also conducted simulations (denoted as Simulation 3) to check the performance of the proposed procedure when the signal is not sufficient and to investigate the validity of treating an ordinal response as a continuous variable, which is the approach taken when we apply the analysis to real data. All the results point to the good performance of the proposed method and hint at the appropriateness of data analysis reported in the next section.

6 Application of the latent variable model

The World Values Survey (WVS) gathers information from participants around the world on contemporary societal issues such as individuals' attitudes about their work and religious beliefs. The goal of the survey is to enable a cross-national, cross-cultural comparison and surveillance of respondents' core values. Namely, participants' responses help identify what or how social and personal factors affect individuals' core values. For this application, we use data from the India cohort ($n = 759$); our specific aim is to investigate whether respondents' financial situation and attitudes about their job (adjusted for demographic factors) influence their core values, as gauged by the following nine questions:

Y_1: How important is God in your life? (1=not at all, 10=very)

Y_2: Overall, how satisfied or dissatisfied are you with your home life? (1=dissatisfied, 10=very satisfied)

Y_3: All things considered, how satisfied are you with your life as a whole in these days? (1=dissatisfied, 10=very satisfied)

Y_4: How satisfied are you with the financial situation of your household? (1=dissatisfied, 10=very satisfied)

Y_5: Overall, how satisfied or dissatisfied are you with your job? (1= dissatisfied, 10=very satisfied);

Y_6: Individuals should take more responsibility for providing for themselves. (1=agree completely, 10=disagree completely)

Y_7: Competition is good. It simulates people to work hard and develop new ideas. (1=agree completely, 10=disagree completely)

Y_8: In the long run, hard work usually brings about a better life. (1=agree completely, 10=disagree completely)

Y_9: How much pride, if any, do you take in the work that you do? (1=a great deal, 2=some, 3=little, 4=none)

Because the outcomes $Y_1, \cdots, Y_8$ are measured on a scale from 1 to 10 and Y_9 takes values of 1 to 4, we treat the first eight outcomes as continuous variables and the last outcome as ordinal. With nine outcomes, it is reasonable to consider at most nine latent variables $\xi_1, \cdots, \xi_9$ in the proposed model:

$$\begin{aligned} Y_k &= b_k + \sum_{j=1}^{9} \alpha_{kj}\xi_j + \varepsilon_k, \;\; k = 1, \ldots, 8, \\ U_9 &= b_9 + \sum_{j=1}^{9} \alpha_{9j}\xi_j + \varepsilon_9, \\ \xi_k &= \mathbf{Z}'\boldsymbol{\gamma}_k + e_k, \;\; k = 1, \ldots, 9, \end{aligned}$$

where $Y_9 = I(U_9 \leq c_1) + 2I(c_1 < U_9 \leq c_2) + 3I(c_2 < U_9 \leq c_3) + 4I(c_3 < U_9)$ and $\mathbf{Z} = (Z_1, \ldots, Z_6)'$, in which (Z_1, Z_2) = marriage ((1, 0), more than once; (0, 0), only once; (0, 1), never), Z_3 = age, Z_4 = gender (1, male; 0, female); Z_5 = income (1: <12,000 rupees per year; 2: 12,001−18,000; 3: 18,001−24,000; 4: 24,001−30,000; 5: 30,001−36,000; 6: 36,001−48,000; 7: 48,001−60,000; 8: 60,001−90,000; 9: 90,001−120,000; and 10: >120,000); and Z_6 = freedom of decision-making on the job (1, none at all; 10, a great deal). To unify scales of covariates, we standardize the elements in $\mathbf{Z}$ before analysis. For identifiability, the matrix $\boldsymbol{\alpha}$ is assumed to be a lower triangular matrix, with 1's as diagonal entries, $b_9 = 0$ and $\sigma_{\varepsilon 9} = 1$. The tuning parameter $\rho_{1n} = 0.2$ and $\rho_{2n} = 0.1$ are chosen by maximizing equations (3.16) and (3.17). We also consider the method without selection of the latent variables and the predictor variables (Non-p); Tables 4−6 display point estimates and the estimated SDs (in parenthesis). We used 1000 Monte Carlo replications to approximate conditional means. We calculated the SDs via the resampling method described in Section 4, with 1000 replications. We decided on a sample size of 1000 by monitoring the stability of the SDs; we found that when the bootstrap sample size was between 500 and 1000, the resulting SDs stabilized and the difference was only marginal. For the proposed method, the algorithm failed to converge in 76 of the 1000 replications; the results from the proposed method are based on 924 replications. The Non-p method

did not fit the data properly, resulting in about 665 of 1000 runs failing to converge; the results from the Non-p method are based on 335 replications. Hence, given the low number of replications, the SDs of the Non-p estimator displayed in Tables 4−6 are likely not representative of the SDs that would result from 1000 runs.

Table 4: Estimates of $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_9$ for WVS data

	$\boldsymbol{\gamma}_1$		$\boldsymbol{\gamma}_2$		$\boldsymbol{\gamma}_3$	
	Proposed	Non-p	Proposed	Non-p	Proposed	Non-p
Z_1	0	-0.130(0.121)	0	-0.136(0.090)	0	0.143(0.087)
Z_2	0	0.060(0.090)	0	-0.032(0.079)	0	0.018(0.065)
Z_3	0	0.143(0.086)	0	0.041(0.076)	0	0.010(0.078)
Z_4	0	-0.104(0.093)	0	0.096(0.081)	0	-0.121(0.076)
Z_5	0	-0.207(0.101)	0.546(0.070)	0.567(0.078)	0	-0.073(0.138)
Z_6	0	0.390(0.096)	0	0.220(0.097)	0	0.066(0.097)
	$\boldsymbol{\gamma}_4$		$\boldsymbol{\gamma}_5$		$\boldsymbol{\gamma}_6$	
	Proposed	Non-p	Proposed	Non-p	Proposed	Non-p
Z_1	0	0.094(0.092)	0	0.022(0.096)	0	-0.109(0.124)
Z_2	0	-0.039(0.073)	0	-0.061(0.063)	0	-0.119(0.119)
Z_3	0	-0.014(0.063)	0	0.079(0.075)	0	0.129(0.106)
Z_4	0	-0.090(0.064)	0	0.046(0.086)	0	-0.137(0.114)
Z_5	0	0.257(0.107)	0	-0.009(0.111)	0	0.176(0.177)
Z_6	0	0.023(0.080)	0.511(0.082)	0.485(0.096)	0	0.113(0.149)
	$\boldsymbol{\gamma}_7$		$\boldsymbol{\gamma}_8$		$\boldsymbol{\gamma}_9$	
	Proposed	Non-p	Proposed	Non-p	Proposed	Non-p
Z_1	0	-0.233(0.112)	0	-0.080(0.173)	0	0.032(0.138)
Z_2	0	-0.085(0.092)	0	-0.087(0.133)	0	0.146(0.099)
Z_3	0	-0.051(0.106)	0	0.110(0.116)	0	0.021(0.115)
Z_4	0	-0.093(0.128)	0	-0.076(0.125)	0	0.060(0.104)
Z_5	0	-0.240(0.162)	0	0.007(0.178)	0	-0.345(0.150)
Z_6	0	-0.167(0.140)	0	-0.039(0.145)	0	0.046(0.170)

Our penalized method enables the estimates of $\|\boldsymbol{\gamma}_j\|$, $j = 1, 3, 4, 6, 7, 8, 9$ (see Table 4), σ_{e4}, and σ_{e9} (see Table 5) to be exactly zero. As discussed in Section 2, $\{\sigma_{e4} = 0, \|\boldsymbol{\gamma}_4\| = 0\}$ and $\{\sigma_{e9} = 0, \|\boldsymbol{\gamma}_9\| = 0\}$ imply that ξ_4 and ξ_9 are zero and can be ignored completely; $\{\sigma_{ej} \neq 0, \|\boldsymbol{\gamma}_j\| = 0,\ j = 1, 3, 6, 7, 8\}$ imply that $\xi_j, j = 1, 3, 6, 7, 8$ are simply random effects; $\{\sigma_{e2} \neq 0, \|\boldsymbol{\gamma}_2\| \neq 0\}$ and $\{\sigma_{e5} \neq 0, \|\boldsymbol{\gamma}_5\| \neq 0\}$ imply that ξ_2 and ξ_5 are indeed latent variables, characterized by income and job freedom

separately.

Although model (3.4) reveals that the dependence among the nine outcomes is explained jointly by random effects and latent variables $\{\xi_2, \xi_5\}$, the two latent constructs add more new insights. First, the significantly positive estimates of factor loadings $\widehat{\alpha}_{32} = 0.896$ (0.148), $\widehat{\alpha}_{42} = 1.015$ (0.144) and $\widehat{\alpha}_{52} = 0.609$ (0.129) (see Table 6) imply that respondents' income level has positive effects on outcomes Y_3, Y_4 and Y_5 (life, finance and job satisfaction). Second, the significantly negative estimates of factor loadings $\widehat{\alpha}_{92} = -0.253$ (0.093) and $\widehat{\alpha}_{95} = -0.486$ (0.125) (see Table 6) reveal that both income and job freedom have positive effects on respondents' feelings of pride in their job, given a reversed coding of Y_9. Third, insignificant factor loadings $\{\alpha_{62}, \alpha_{72}, \alpha_{82}\}$ and $\{\alpha_{65}, \alpha_{75}, \alpha_{85}\}$ (see Table 6) indicate negligible effects of income and job freedom on outcomes Y_6, Y_7 and Y_8 (sense of responsibility, competitiveness and work intensity). Finally, the two latent constructs help in the interpretation of the heterogeneities among subjects. For example, people with similar levels of income and perceived job freedom tend to give similar answers to questions Y_3, Y_4, Y_5 and Y_9 (life, finance, job satisfaction and pride in work). In summary, our results render statistical evidence for some well-known but hard-to-measure social psychology phenomena.

Table 5: Estimates of c and variance for WVS data

	Proposed(SD)	Non-p(SD)		Proposed(SD)	Non-p(SD)
c_1	0.140(0.053)	0.299(0.107)	c_2	1.506(0.082)	3.209(0.662)
c_3	2.624(0.156)	5.590(1.133)			
$\sigma^2_{\varepsilon 1}$	4.487(0.951)	3.875(0.548)	σ^2_{e1}	2.025(0.860)	2.368(0.523)
$\sigma^2_{\varepsilon 2}$	1.515(0.468)	0.902(0.539)	σ^2_{e2}	1.368(0.462)	2.150(0.668)
$\sigma^2_{\varepsilon 3}$	1.979(0.323)	1.307(0.614)	σ^2_{e3}	0.705(0.345)	1.598(0.905)
$\sigma^2_{\varepsilon 4}$	2.135(0.432)	1.731(0.694)	σ^2_{e4}	0	0.780(0.848)
$\sigma^2_{\varepsilon 5}$	1.907(0.863)	1.360(0.775)	σ^2_{e5}	1.052(0.904)	1.652(0.888)
$\sigma^2_{\varepsilon 6}$	3.760(2.330)	3.815(1.880)	σ^2_{e6}	3.120(2.336)	2.995(1.998)
$\sigma^2_{\varepsilon 7}$	1.981(1.188)	2.224(0.899)	σ^2_{e7}	2.054(1.534)	2.004(1.313)
$\sigma^2_{\varepsilon 8}$	3.221(1.732)	3.478(1.223)	σ^2_{e8}	0.603(1.966)	0.617(1.644)
$\sigma^2_{\varepsilon 9}$	1	1	σ^2_{e9}	0	3.587(1.591)

Unlike ordinary multiple regression models, which account for the effects of covariates on outcomes separately, the general LVM proposed in this study groups multiple outcomes into two latent constructs, which reduces the model dimension, simultane-

ously accommodates dependence between outcomes and heterogeneity between subjects and provides simpler interpretation of the associations among multidimensional outcomes.

Table 6: Estimates of $\boldsymbol{\alpha}$ for WVS data

	Proposed(SD)	Non-p(SD)		Proposed(SD)	Non-p(SD)
α_{21}	0.742(0.316)	0.540(0.114)	α_{42}	1.015(0.144)	0.627(0.192)
α_{31}	0.581(0.278)	0.351(0.123)	α_{52}	0.609(0.129)	0.430(0.136)
α_{41}	0.369(0.217)	0.312(0.098)	α_{62}	0.205(0.170)	-0.030(0.108)
α_{51}	0.385(0.163)	0.310(0.103)	α_{72}	-0.187(0.128)	-0.137(0.099)
α_{61}	0.039(0.141)	0.126(0.108)	α_{82}	0.128(0.146)	0.013(0.108)
α_{71}	-0.218(0.123)	-0.178(0.100)	α_{92}	-0.253(0.093)	-0.320(0.091)
α_{81}	-0.447(0.148)	-0.358(0.112)	α_{43}	0.739(0.350)	0.531(0.149)
α_{91}	-0.237(0.079)	-0.463(0.098)	α_{53}	0.598(0.353)	0.362(0.190)
α_{32}	0.896(0.148)	0.750(0.208)	α_{63}	-0.369(0.367)	-0.108(0.196)
α_{73}	0.447(0.331)	0.094(0.175)	α_{75}	-0.228(0.149)	-0.082(0.193)
α_{83}	-0.504(0.367)	-0.295(0.258)	α_{85}	-0.227(0.186)	-0.091(0.210)
α_{93}	-0.272(0.179)	-0.412(0.155)	α_{95}	-0.486(0.125)	-0.770(0.222)
α_{54}	0	0.386(0.292)	α_{76}	0.500(0.299)	0.456(0.330)
α_{64}	0	0.227(0.456)	α_{86}	0.299(0.284)	0.337(0.295)
α_{74}	0	0.244(0.442)	α_{96}	0.030(0.052)	0.060(0.113)
α_{84}	0	0.271(0.417)	α_{87}	0.853(0.269)	0.762(0.281)
α_{94}	0	-0.047(0.249)	α_{97}	0.063(0.067)	0.154(0.142)
α_{65}	0.048(0.153)	-0.101(0.190)	α_{98}	-0.123(0.140)	0.003(0.290)

7 Discussion

We have proposed a penalized ML estimator to develop a general framework of latent variable selection. The proposed method is able to select latent variables and estimate parameters simultaneously. Under mild conditions, the estimator is $n^{1/2}$-consistent and asymptotically normal. Given an appropriate choice of regularization parameters, the proposed estimator demonstrates the oracle property. We suggest using a BIC-type tuning parameter selection method to select the regular parameters.

We have focused on mixed outcomes with ordinal and continuous variables under the linear regression framework. Because the assumption of normality may not always

be practical, our future work will extend our methods to other regression frameworks (e.g., generalized linear regression) for non-normal responses. Moreover, we have focused on selecting important latent variables, but one can easily extend the proposed method to simultaneously select manifest variables and latent variables.

Acknowledgements

Huazhen Lin's research is supported by the National Natural Science Foundation of China (Nos. 11071197 and 11125104) and the Program for New Century Excellent Talents at the University of China. Xin-Yuan Song's research is supported by a grant from the Research Grant Council of the Hong Kong Special Administration Region (GRF 404711). Yi Li's research is partially supported by a grant from the U.S. National Institutes of Health (grant no. 1R01HL1HL107240). We thank the Editor, an Associate Editor and two referees for their insightful suggestions, which helped significantly improve this manuscript.

Supporting Information

Additional information for this article is available online including:

Supplementary material contains detailed proofs of Theorems 1-4, Figures 1-5, Simulations 2-3 and Tables 2-3.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, **52**, 317–332.

Anderson, T. W. & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 110–150. University of California Press, Berkeley.

Bhattacharya, A. & Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Carvalho, C., Lucas, J., Wang, Q., Nevins, J. & West, M. (2005). High-dimensional sparse factor models and latent factor regression. *ISDS Discussion Paper 2005-15*, Duke University.

Catalano, P. J. & Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Statist. Assoc.*, **87**, 651–658.

Chen, Z. & Dunson, D. B.(2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769.

Cox, D. R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441–461.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *J. R. Statist. Soc. B.*, **62**, 355–366.

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, **7**, 551–568.

Fan, J. Q. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, **96**, 1348–1360.

Fan, J., Lin, H. Z. & Zhou, Y. (2006). Local partial likelihood estimation for life time data. *Ann. Statist.*, **34**, 290–325.

Fitzmaurice, G. M. & Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *J. Am. Statist. Assoc.*, **90**, 845–852.

Gueorguieva, R. V. & Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Statist. Assoc.*, **96**, 1102–1112.

Jin, Z., Ying, Z. & Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, **88**, 381–390.

Lee, S. Y. (2007). *Structural equation modelling: a Bayesian approach.* John Wiley & Sons, Inc.

Lee, S. Y. & Song, X. Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, **29**, 23–39.

Lee, S. Y. & Song, X. Y. (2004). Maximum likelihood analysis of a general latent variable model with hierarchically mixed data. *Biometrics*, **60**, 624–636.

Leek, J. T. & Storey, J. D. (2008). A general framework for the multiple testing dependence. *Proc. Natl. Acad. Sci.*, **105**, 18718–18723.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorial and continuous latent variable indicators. *Psychometrika*, **49**, 115–132.

Olsen, M. K. & Schafer, J. (2001) A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association*, **96**, 730-745.

Press, S. J. & Shigemasu, K. (1989). Bayesian inference in factor analysis. *Contributions to Probability and Statistics:Essays in Honor of Ingram Olkin,* Gleser, L.J., Perlman, M. D., S. J. Press & A.R. Sampson (eds.), Springer Verlag, New York, 271–287.

Press, S. J. & Shigemasu, K. (1997). Bayesian inference in factor analysis-revised, with an appendix by Daniel B. Rowe, Technical Report No. 243, Department of Statistics, University of California, Riverside, CA, May 1997.

Regan, M. M. & Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics*, **55**, 760–768.

Roy, J. & Lin, X. H. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, **56**, 1047–1054.

Sammel, M. D., Ryan, L. M. & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc. B.*, **59**, 667–678.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shi, J. Q. & Lee, S. Y. (2000). Latent variable models with mixed continuous and polytomous data. *J. R. Statist. Soc. B.*, **62**, 77–87.

Skrondal, A. & Hesketh S. R. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, **72**, 123–140.

Skrondal, A. & Rabe-Hesketh, S. (2007). Latent variable modeling: A survey. *Scand. J. Statist.*, **34**, 712–745.

Wagner, H. & Tüchler, R. (2010). Bayesian estimation of random effects models for multivariate responses of mixed data. *Comp. Statist. Data Anal.*, **54**, 1206–1218.

Wang, H., Li, R. & Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

Yang, Y., Jian, K. & Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Stat. Med.*, **26**, 3782–3800.

Zhang, Y., Li, R. & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *J. Am. Statist. Assoc.*, **105**, 312–323.

Huazhen Lin, Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China.
E-mail: linhz@swufe.edu.cn

Appendix

A.1 Notation

Let the parameter $\mathbf{\Theta} = (\mathbf{\Theta}_1', \mathbf{\Theta}_2', \mathbf{\Theta}_3')'$, where $\mathbf{\Theta}_1 = (\vec{\boldsymbol{\alpha}}', \vec{\boldsymbol{\beta}}', \boldsymbol{\sigma}_\varepsilon')'$, $\mathbf{\Theta}_2 = (\Theta_{21}, \cdots, \Theta_{2q})' = \boldsymbol{\sigma}_e$ and $\mathbf{\Theta}_3 = (\Theta_{31}, \cdots, \Theta_{3,q\times m})' = \vec{\boldsymbol{\gamma}}$, m is the length of $\mathbf{Z}_i$. Let threshold $c_j(y) = c_{j,y}$, $c_{j0}(y) = c_{j,y0}$,

$$d_{kj}(y) = \frac{E\phi\left(\frac{c_{j0}(y)-W_{ij}(\mathbf{\Theta})}{\nu_j}\right)\left\{\frac{\partial W_{ij}(\mathbf{\Theta})}{\partial \mathbf{\Theta}_k} + [c_{j0}(y) - W_{ij}(\mathbf{\Theta})]\frac{\partial \log(\nu_j)}{\partial \mathbf{\Theta}_k}\right\}}{E\phi\left(\frac{c_{j0}(y)-W_{ij}(\mathbf{\Theta})}{\nu_j}\right)}|_{\mathbf{\Theta}=\mathbf{\Theta}_0},$$

where $d_{kj}(y)$ is the derivative of $\widehat{c}_j(\mathbf{\Theta}; y)$ with respect to $\mathbf{\Theta}_k$ at $\mathbf{\Theta} = \mathbf{\Theta}_0$, $\widehat{c}_j(\mathbf{\Theta}; y)$ is the estimator of $c_j(y)$ given $\mathbf{\Theta}$, $\nu_j = \sqrt{\boldsymbol{\alpha}_j'\mathbf{\Sigma}_e\boldsymbol{\alpha}_j + 1}$, and $W_{ij}(\mathbf{\Theta}) = \mathbf{X}_i'\beta_j + \boldsymbol{\alpha}_j'\boldsymbol{\gamma}\mathbf{Z}_i$.

Similar to $\boldsymbol{\sigma}_e = (\boldsymbol{\sigma}_{e(1)}', \boldsymbol{\sigma}_{e(2)}')'$ or $\vec{\boldsymbol{\gamma}} = (\vec{\boldsymbol{\gamma}}_{(1)}', \vec{\boldsymbol{\gamma}}_{(2)}')'$, let $\mathbf{\Theta}_2 = (\Theta_{2(1)}', \Theta_{2(2)}')'$, $\mathbf{\Theta}_3 = (\Theta_{3(1)}', \Theta_{3(2)}')'$, $d_{2j}(y) = (d_{2j(1)}(y)', d_{2j(2)}(y)')'$ and $d_{3j}(y) = (d_{3j(1)}(y)', d_{3j(2)}(y)')'$. Let

$$\begin{aligned}
B_{(rs)} = E\Bigg(&\frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\Theta_{r(1)}\partial\Theta_{s(1)}'} \\
&+ \sum_{j=p_1+1}^{p}\left(\frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\Theta_{r(1)}\partial c_j(Y_{ij})}d_{sj(1)}'(Y_{ij}) + \frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\Theta_{r(1)}\partial c_j(Y_{ij}-1)}d_{sj(1)}'(Y_{ij}-1)\right)\Bigg) \text{ and} \\
B_{rs} = E\Bigg(&\frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\mathbf{\Theta}_r\partial\mathbf{\Theta}_s'} \\
&+ \sum_{j=p_1+1}^{p}\left(\frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\mathbf{\Theta}_r\partial c_j(Y_{ij})}d_{sj}'(Y_{ij}) + \frac{\partial^2 logL_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial\mathbf{\Theta}_r\partial c_j(Y_{ij}-1)}d_{sj}'(Y_{ij}-1)\right)\Bigg), \quad \text{(A.1)}
\end{aligned}$$

where $L_i(\mathbf{\Theta}; \mathbf{c})$ is the observed-data likelihood function for subject i. The matrix $B = (B_{rs})$ is the mean of the Hessian matrix of $\log L_n(\mathbf{\Theta}; \widehat{\mathbf{c}}(\mathbf{\Theta}))$ respect to $\mathbf{\Theta}$ and the matrix $(B_{(rs)})$ is B corresponding to non-zero components of $\mathbf{\Theta}$.

Let

$$\begin{aligned}
&\mathcal{U}_1 = \text{diag}\{\ddot{p}_{\rho_{1n}}(\sigma_{e10}), \ldots, \ddot{p}_{\rho_{1n}}(\sigma_{es0})\}; \;\; \mathbf{b_1} = \left(\dot{p}_{\rho_{1n}}(\sigma_{e10}), \ldots, \dot{p}_{\rho_{1n}}(\sigma_{es0})\right)', \\
&\mathcal{U}_2 = \text{diag}\left\{\ddot{p}_{\rho_{2n}}(|\gamma_{110}|), \cdots, \ddot{p}_{\rho_{2n}}(|\gamma_{1,h_1,0}|), \cdots, \ddot{p}_{\rho_{2n}}(|\gamma_{q,1,0}|), \cdots, \ddot{p}_{\rho_{2n}}(|\gamma_{q,h_q,0}|)\right\}, \\
&\mathbf{b_2} = \big(\dot{p}_{\rho_{2n}}(|\gamma_{110}|)sgn(\gamma_{110}), \ldots, \dot{p}_{\rho_{2n}}(|\gamma_{1,h_1,0}|)sgn(\gamma_{1,h_1,0}), \\
&\qquad\qquad \cdots, \dot{p}_{\rho_{2n}}(|\gamma_{q10}|)sgn(\gamma_{q10}), \ldots \dot{p}_{\rho_{2n}}(|\gamma_{q,h_q,0}|)sgn(\gamma_{q,h_q,0})\big)'.
\end{aligned}$$

$\mathcal{U}_1$ and $\mathcal{U}_2$ are used to express the uncertainty due to adding the penalties on $\mathbf{\Sigma}_e$ and $\boldsymbol{\gamma}$, respectively, while $\mathbf{b_1}$ and $\mathbf{b_2}$ are corresponding biases.

Denote

$$B_{(rs.k)} = B_{(rs)} - B_{(rk)}B_{(kk)}^{-1}B_{(ks)},\ \ B^*_{(rs.k)} = B_{(rs)} - B_{(rk)}\left(B_{(kk)} - \mathcal{U}_1\right)^{-1}B_{(ks)},$$

$$\mathbf{\Lambda}_1 = -B^*_{(11.2)} + B^*_{(13.2)}\left(B^*_{(33.2)} - \mathcal{U}_2\right)^{-1}B^*_{(31.2)},$$
$$\mathbf{\Lambda}_2 = -B_{(22.1)} + B_{(23.1)}\left(B_{(33.1)} - \mathcal{U}_2\right)^{-1}B_{(32.1)},$$
$$\mathbf{\Lambda}_3 = -B_{(33.1)} + B_{(32.1)}\left(B_{(22.1)} - \mathcal{U}_1\right)^{-1}B_{(23.1)}.$$

$A_k, k = 1, 2, 3$ and $A_{4j}(y)$ are defined as:

$$A_k = E\left[\left(m_{k1}\Upsilon_{i(1)} + m_{k2}\Upsilon_{i(2)} + m_{k3}\Upsilon_{i(3)}\right)^{\otimes 2}\right],$$
$$A_{4j}(y) = E\left[\left(\frac{\nu_{j0}}{\psi_j(y)}\varpi_{ij}(y) - \left(m_{4j1}(y)\Upsilon_{i(1)} + m_{4j2}(y)\Upsilon_{i(2)} + m_{4j3}(y)\Upsilon_{i(3)}\right)\right)^{\otimes 2}\right], \quad \text{(A.2)}$$

where $\mathbf{\Lambda}_1^{-1}A_1(\mathbf{\Lambda}_1')^{-1}$, $\mathbf{\Lambda}_2^{-1}A_2(\mathbf{\Lambda}_2')^{-1}$, $\mathbf{\Lambda}_3^{-1}A_3(\mathbf{\Lambda}_3')^{-1}$ and $A_{4j}(y)$ are asymptotic standard errors of $\sqrt{n}(\widehat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_{10})$, $\sqrt{n}(\widehat{\boldsymbol{\sigma}}_{e(1)} - \boldsymbol{\sigma}_{e(1)0})$, $\sqrt{n}(\widehat{\vec{\gamma}}_{(1)} - \vec{\gamma}_{(1)0})$ and $\sqrt{n}(\widehat{c}_{j,y} - c_{j,y0})$, respectively, when zero components are known, and

$$m_{11} = m_{22} = m_{33} = -1,\ \ m_{13} = B^*_{(13.2)}\left(B_{(22)} - \mathcal{U}_1\right)^{-1},$$

$$m_{12} = -\left({B_{(13.2)}}^*\left(B^*_{(33.2)} - \mathcal{U}_2\right)^{-1}B_{(32)} - B_{(12)}\right)\left(B_{(22)} - \mathcal{U}_1\right)^{-1},$$

$$m_{21} = (B_{(21)}{B_{(11)}}^{-1} - B_{(23.1)}\left(B_{(33.1)} - \mathcal{U}_2\right)^{-1}B_{(31)}{B_{(11)}}^{-1}),$$

$$m_{23} = B_{(23.1)}\left(B_{(33.1)} - \mathcal{U}_2\right)^{-1},\ \ m_{32} = B_{(32.1)}\left(B_{(22.1)} - \mathcal{U}_1\right)^{-1},$$

$$m_{31} = (B_{(31)} - B_{(32.1)}\left(B_{(22.1)} - \mathcal{U}_1\right)^{-1}B_{(21)}){B_{(11)}}^{-1},$$

$$m_{4jk}(y) = d_{1j}(y)'\mathbf{\Lambda}_1^{-1}m_{1k} + d_{2j(1)}(y)'(\mathbf{\Lambda}_2 + \mathcal{U}_1)^{-1}m_{2k} + d_{3j(1)}(y)'(\mathbf{\Lambda}_3 + \mathcal{U}_2)^{-1}m_{3k},$$

$$\Upsilon_{i(k)} = \frac{\partial log L_i\left(\mathbf{\Theta}_0; \mathbf{c}_0\right)}{\partial \Theta_{k(1)}} + \sum_{j=p_1+1}^{p}\left(\varphi_{ij1,(k)} + \varphi_{ij2,(k)}\right),$$

$$\varphi_{rj1,k} = E\left\{\frac{\partial^2 log L_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial \mathbf{\Theta}_k \partial c_j(Y_{ij})}\frac{\nu_{j0}}{\psi_j(Y_{ij})}\varpi_{rj}(Y_{ij})|\mathbf{Y}_r, \mathbf{X}_r, \mathbf{Z}_r\right\},$$

$$\varphi_{rj2,k} = E\left\{\frac{\partial^2 log L_i(\mathbf{\Theta}_0; \mathbf{c}_0)}{\partial \mathbf{\Theta}_k \partial c_j(Y_{ij}-1)}\frac{\nu_{j0}}{\psi_j(Y_{ij}-1)}\varpi_{rj}(Y_{ij}-1)|\mathbf{Y}_r, \mathbf{X}_r, \mathbf{Z}_r\right\},$$

$$\psi_j(y) = E\phi\left(\frac{c_{j0}(y) - W_{ij}(\mathbf{\Theta}_0)}{\nu_{j0}}\right),\ \ \varpi_{ij}(y) = I(Y_{ij} \le y) - \Phi\left(\frac{c_{j0}(y) - W_{ij}(\mathbf{\Theta}_0)}{\nu_{j0}}\right),$$

where ν_{j0} is the true value of ν_j, $\varphi_{rj1,(k)}$ and $\varphi_{rj2,(k)}$ are the corresponding parts of $\varphi_{rj1,k}$ and $\varphi_{rj2,k}$ to non-zero parameters, respectively.

Finally, let $C_{21} = C_{32} = 1$, $C_{12} = -B^*_{(13.2)} \left(B^*_{(33.2)} - \mathcal{U}_2 \right)^{-1}$,

$$C_{11} = - \left(B_{(12)} - B^*_{(13.2)} \left(B^*_{(33.2)} - \mathcal{U}_2 \right)^{-1} B_{(32)} \right) (B_{(22)} - \mathcal{U}_1)^{-1},$$

$$C_{22} = -B_{(23.1)} \left(B_{(33.1)} - \mathcal{U}_2 \right)^{-1}, \; C_{31} = -B_{(32.1)} \left(B_{(22.1)} - \mathcal{U}_1 \right)^{-1} \text{ and}$$

$$C_{4jk}(y) = -d_{1j}(y)' \mathbf{\Lambda}_1^{-1} C_{1k} - d_{2j(1)}(y)' (\mathbf{\Lambda}_2 + \mathcal{U}_1)^{-1} C_{2k} - d_{3j(1)}(y)' (\mathbf{\Lambda}_3 + \mathcal{U}_2)^{-1} C_{3k}.$$

A.2 Conditions

(1) The matrix $B = (B_{rs})_{r,s=1,2,3}$ defined by equation (A.1) is negative definite.

(2) A_1, A_2, A_3 and $A_{4j}(k)$ defined by equation (A.2) are positive definite matrices.

(3) $\mathbf{X}_i$ and $\mathbf{Z}_i$ are bounded.

Condition (1) is an identifiability condition for $\mathbf{\Theta}$. A_1, A_2, A_3 and $A_{4j}(k)$ are asymptotic variances of $\sqrt{n}\mathbf{\Lambda}_1(\widehat{\mathbf{\Theta}}_1 - \mathbf{\Theta}_{10})$, $\sqrt{n}\mathbf{\Lambda}_2(\widehat{\boldsymbol{\sigma}}_{e(1)} - \boldsymbol{\sigma}_{e(1)0})$, $\sqrt{n}\mathbf{\Lambda}_3(\widehat{\vec{\gamma}}_{(1)} - \vec{\gamma}_{(1)0})$ and $\sqrt{n}(\widehat{c}_{j,k} - c_{j,k0})$, respectively.